Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 69 (2015) 55 – 65

Procedia Computer Science

7th International Conference on Advances in Information Technology

Enhancing Reliability through Screening and Segmentation: An Online Video Subjective Quality of Experience Case Study

Mark Chignell[a], Weiwei Li[a], Alberto Leon-Garcia[a], Leon Zucherman[b], and Jie Jiang[b]

[a]*University of Toronto, 27 King's College Circle, Toronto, M5S 1A1, Canada*
[b]*TELUS Communications Company, 2455 Cawthra Road #55, Mississauga, L5A 3P1, Canada*

Abstract

In this paper we examine the reliability of subjective rating judgments along a single dimension, focusing on estimates of technical quality produced by integrity impairments and failures (non-accessibility, and non-retainability) associated with viewing video. There is often considerable variability, both within and between individuals, in subjective rating tasks. In the research reported here we consider different approaches to screening out unreliable participants. We review available alternatives, including a method developed by the ITU, a method based on screening outliers, a method based on strength of correlations with an assumed "natural" ordering of impairments, and a clustering technique that makes no assumptions about the data. We report on an experiment that assesses subjective quality of experience associated with impairments and failures of online video. We then assess the reliability of the results using a correlation method and a clustering method, both of which give similar results. Since the clustering method utilized here makes fewer assumptions about the data, it may be a useful supplement to existing techniques for assessing reliability of participants when making subjective evaluations of the technical quality of videos.

Peer-review under responsibility of the Organizing Committee of IAIT2015

Keywords: Reliability; Quality of Experience; Subjective Assessment; Cluster Analysis

1. Introduction

Monitoring and control of quality is an important aspect of many services. In some cases, quality of experience may be predicted algorithmically (e.g., [1]). However, in the case of video impairments it is not clear that an algorithmic prediction of quality of experience is feasible. For instance, people might feel that one or two instances of freezing of the video are ok, but perceive a large decrement in experienced quality if further instances of freezing

Peer-review under responsibility of the Organizing Committee of IAIT2015
doi:10.1016/j.procs.2015.10.006

occur. Since there are no obvious algorithms for predicting quality of experience in the face of video impairments it seems natural to use subjective ratings such as the mean opinion score (MOS) developed by the ITU[2] and associated researchers.

Since individual subjective ratings are subject to error, the judgments of a number of participants are typically averaged to obtain estimates of the "true" values of the construct being judged. However, there may be participants who are unmotivated, incapable of judging the construct accurately, or whose judgments may be unreliable (e.g., they are making judgments on two dimensions rather than an assumed single dimension).

In this paper we address the issue of how participants should be screened when subjectively rating aspects of services such as online video. We focus in particular on the task of subjectively rating quality of experience for videos that have impairments and failures. After reviewing a number of approaches we report on an experiment where participants rated the technical quality of online videos that had associated impairments and failures. We present a case study on using correlational, and cluster, analysis to identify unreliable participants within a sample.

2. Background

The literature on scaling in psychometrics and psychophysics has been dominated by three main tasks, namely judgments of intensity or magnitude, judgments of proximity or similarity, and hedonic (preference, or liking) judgments. Depending on one's perspective subject quality of experience (SQE) judgments can be seen as involving intensity (e.g., what was the overall quality of the experience, what was the technical quality of the experience) and/or hedonic (e.g., how much did I enjoy the experience, what is my preference for the experience vs. other experiences) components. [3] provides a relatively early review of intensity scaling methods, while [4]'s review reflects more recent interest in hedonic scaling, and [5] provide a review of multidimensional scaling of proximities, similarities and preferences.

Research on SQE (see [6] for a recent review) has generally assumed that it is possible for people to give relatively accurate ratings of their experience. As noted by [7], video quality is usually measured using a five-point scale, where a score of 1 means lowest video quality and a score of 5 means highest video quality. The justification [4]for treating humans like measuring rulers is that it often seems to work. For instance, [8] found that absolute ratings of videos presented one at a time produced "repeatable subjective results, even across different scales and different groups of participants."

However, humans are not always perfect measurement rulers and the question then arises of how to deal with inconsistencies in rating. There have been a number of attempts to deal with such inconsistencies in subjective quality of experience experiments. [9] discussed methods for dealing with inconsistency. One frequently used approach is to remove statistical outliers. This can be done on a per-trial basis or at the level of the study participant. The idea is to characterize a collection of results as a distribution (typically a normal distribution) and then to remove results as outliers if they are in the tail of the distribution (e.g., at a percentile of 97.5%, or 99%, or greater). The removal of statistical outliers can be problematic because it doesn't take into account the accuracy or consistency of judgment, but simply removes trials or participants based on statistical departures from average performance.

Another approach is to explicitly model participants as being "reliable" or "unreliable". [9] examined the issue of reliability in a challenging observational setting where participants made judgments of SQE in their own ("crowdsourced") environment (i.e., in the absence of a supervising experimenter, with relatively anonymous participants, and with no control over where the participant chooses to carry out the experiment). They asked screening questions to check if the participant was paying attention. An example of a content related screening question was "which of the following animals did you see in the video", and a quality related question example was "did you notice any stops in the video that you just watched?".

Another method for assessing reliability cited by [9] included identifying participants as reliable if their judgments had relatively high correlations with mean scores for the entire sample participants. In this paper we develop a clustering approach to differentiating participants that is distribution free and that can be carried out automatically so as to remove the possibility of bias. Our ultimate goal is to identify likely causes of unreliability and to develop a method for screening participants so as to increase experimental efficiency. We compare our clustering approach

with a modified version of the correlational approach that takes advantage of the expected ordering of technical quality scores across impairments with different numbers of freezings.

3. Experiment

A total of 21 undergraduate and graduate students were recruited from the University of Toronto via listservs, poster advertisements, word of mouth and class announcements. The inclusion criteria included the following; over 18 years of age, normal visual acuity (corrective and lenses are allowed). The exclusion criteria were the following; participated in a video quality experiment in the past six months and/or has no experience or interest in watching video through the Internet.

The experiment was performed at the University of Toronto, according to an experimental protocol approved by the University's ethics review board. Participants performed the experiment one at a time. A series of short videos (each around 1-2 minutes in length) were viewed and judgments of acceptability (yes/no), technical quality, content quality, and overall quality of experience were collected. The latter three ratings were collected using a standard five-point MOS (mean opinion score) scale with five rating anchors (excellent, good, fair, poor, bad). Of the 21 participants tested, one participant decided to leave during the test, and this participant's incomplete scoring was not used in the reliability analysis reported below.

Table 1 shows the impairment types tested in the experiment. We used a set of 30 original unimpaired videos and then added impairments and failures to those videos to create the instances of impairments and failures used in the experiment. The Integrity impaired video had between 1 and 4 interruption events of duration 10 seconds each as shown in Table 1. The videos with non-Retainability failures had premature endings, at either the 20 second or 40 second mark of the video. The videos with non-Accessibility failures displayed a failure-to-play message either immediately or after 10 seconds. In both cases, none of the video was actually seen. Each participant saw the same mixture of pristine videos, impairments and failures, and the same set of 30 videos, but different participants saw different mappings, of impairments and failures, to videos.

Table 1. Description of Video Impairments and Failures.

Impairment and Failures	Description of Impairment
H0	Unimpaired (Pristine)
H1	Single temporary interruption of 10s duration happening at 40s (time after video playback start)
H2	Two 10s (temporary) interruptions happening at 20s and 40s respectively
H3	Three 10s (temporary) interruptions happening at 10s, 20s and 40s respectively
H4	Four 10s (temporary) interruptions happening at 10s, 20s, 30s, and 40s respectively
NR1	A permanent interruption happening at 20s
NR2	A permanent interruption happening at 40s
NA1	Video never starts to play. Video player displays "failure-to-play" message immediately
NA2	Video never starts to play. Video player displays "failure-to-play" message after 10s

The experiment was divided into three blocks, each involving assessment of 10 videos. In Block 1, a subject viewed a mix of pristine videos and videos with Integrity impairments. In Blocks 2 and 3, videos with non-Accessibility and non-Retainability failures were added to the mix of pristine and Integrity-impaired videos.

After signing a consent form, the participants filled out an online pre-questionnaire consisting of demographic questions, and questions related to video viewing habits and self-assessed levels of patience and tolerance to

frustration. Next, the subjects viewed the 30 videos in the three blocks described above. The subjects viewed the videos on an LCD screen and input their responses using a keyboard and mouse. The Absolute Category Rating (ACR) method recommended in [10] was used. After viewing each video, subjects answered questions about the viewing experience (these questions are shown in Table 2).

The (spatial) resolution of each video was 512x288 pixels. The frame-rate of each video was 30 frames-per-second. The videos consisted of clips of different lengths that varied between 56 seconds to 123 seconds. Of the 30 video clips, 22 were movie trailers of short duration (teaser-trailers) and 8 were short movies.

The MOS (mean opinion score) value used in the experiment corresponded to session MOS or session Mean Opinion Score (sMOS). Session MOS extends the concept of Mean Opinion Score (MOS) by adding the effect of Session Failures (that can occur with regard to non-Accessibility and non-Retainability failures) to the effect of Session Impairments (traditionally evaluated by MOS). In this case participant experience is determined by any impairment or failure event during the entire life cycle of a service session (viewing of a single video).

4. Results

Average sMOS (across all participants) was calculated for each of the impairments. Fig. 1 shows the result, with error bars indicating 95% confidence intervals around the mean. As expected, sMOS drops with increasing integrity impairment. Across all the participants in the sample, there was no significant difference in mean sMOS ratings between H1 and H2 but there was a large drop between H3 and H4 (comparable to the drop between H0 and H1. Another feature of the data was that NA1 was anchored at the bottom of the scale while NR was rated slightly higher, with a tendency for NR2 to score slightly higher than NR1 (but without reaching statistical significance). The bar chart shows sMOS scores collected across the entire experiment (i.e., the data were pooled across blocks 1, 2, and 3 for each participant).

Table 2. Video Questionnaire.

No.	Rating Criterion/Question	Possible Ratings/Answers
1	Is the technical quality of this video acceptable?	Yes/No
2	Your overall evaluation of the technical quality in the video is:	Excellent/Good/Fair/Poor/Bad
3	The content of the video was:	Very interesting /Interesting /Neutral/Boring/Very boring
4	Your overall viewing experience (Content + Technical Quality) during the video playback was:	Excellent/Good/Fair/Poor/Bad

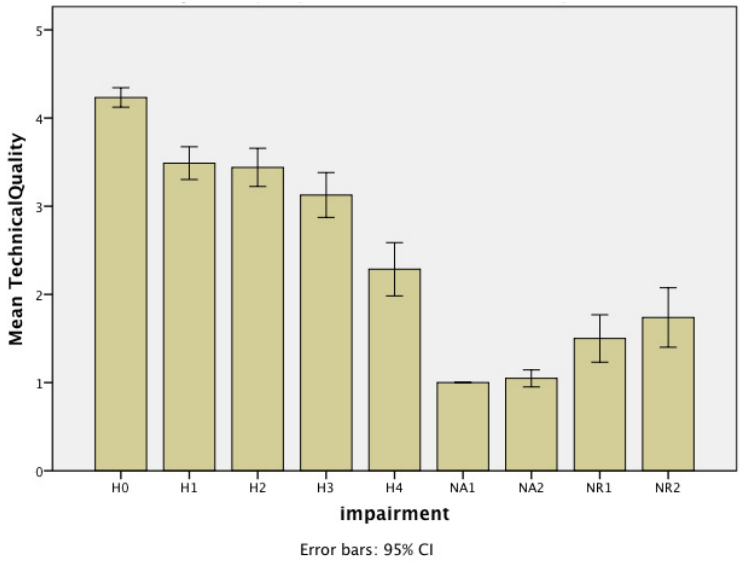

Fig. 1. Mean sMOS by impairment and failure type across all participants.

The correlations between sMOS, and subjective ratings of content quality, and overall quality, were all significant ($p < 0.001$). The correlation between sMOS and overall experience was particularly high ($r = 0.859$), and there was also a fairly strong correlation between sMOS and rated content quality ($r = 0.522$). However, when controlling for overall experience, the partial correlation between sMOS and content was much lower, although still statistically significant ($r = -0.154$). These relationships suggest that it may be appropriate to adjust sMOS scores based on content effects.

4.1. Reliability of subjective judgments

We then carried out a series of analyses aimed at identify "Reliable" participants. Videos with a greater number of freezings were generally judged more harshly than videos with fewer freezings. Figure 2 shows mean ratings for each of the 20 participants who completed the experiment. It can be seen that there is considerable variation in how participants rated impairments and failures. For instance, P040 gave high ratings to impairments and failures, with the exception of non-accessibility failures, which were assigned the lowest rating of 1. P014, and P015, on the other hand showed the expected pattern of results with sMOS decreasing steadily with increasing integrity impairment, and dropping to the floor of the scale for NA and NR impairments. The variability in how responsive participants are to severity of impairment suggests that some participants may be less reliable than others and that an improved estimate of the relationship between sMOS and impairments might be obtained by removing some participants from the analysis, or by segmenting participants into separate groups that have different rating characteristics.

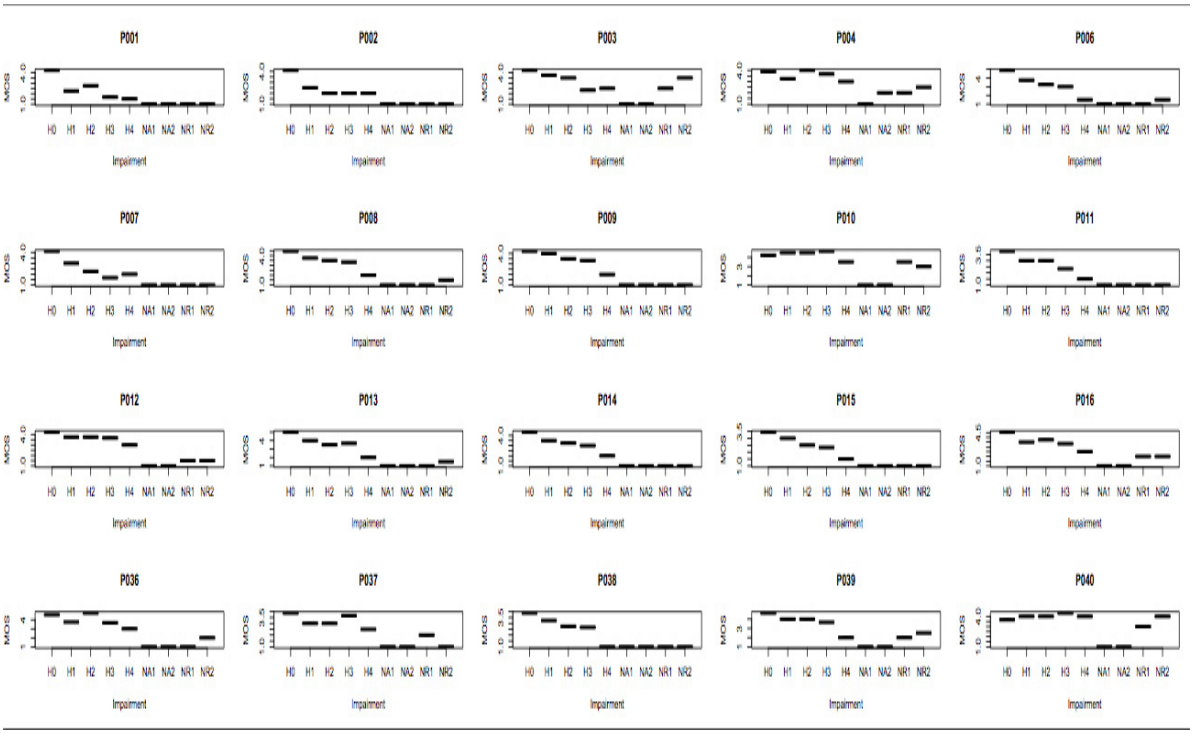

Fig. 2. Mean ratings per participant.

Inspection of the individual data points raised some concerns about how accurate some participants were in making their judgments. Six participants answered "Not-acceptable" to H0 (unimpaired) videos. Other examples of questionable judgment included:

- Scoring 3 (good) for NR scenario
- Scoring 1 (Bad) for H0 (good quality video, no impairments)

The impact of adding failures in blocks 2 and 3 on impairment ratings can be seen in Table 3. The ratings of all the impairments are higher in blocks 2 and 3. One explanation for this effect might be that the addition of the failures filled in the lower parts of the rating scale, forcing the ratings of the less severe impairments upwards .

Table 3. Means and standard errors of impairments across the three blocks.

	Block 1			Block 2			Block 3		
	N	**Mean**	**SE**	**N**	**Mean**	**SE**	**N**	**Mean**	**SE**
H0	63	3.97	0.111	63	4.27	0.079	63	4.46	0.09
H1	42	3.38	0.118	19	3.63	0.232	23	3.57	0.187
H2	42	3.24	0.159	12	3.67	0.31	30	3.63	0.162
H3	42	3.1	0.159	12	3.08	0.288	9	3.33	0.333
H4	21	2	0.183	15	2.73	0.3	6	2.17	0.167
NA1				12	1	0	30	1	0
NA2				34	1.06	0.059	8	1	0
NR1				15	1.53	0.215	27	1.48	0.172
NR2				28	1.79	0.208	14	1.64	0.289

4.2. Identifying Unreliable Participants based on Correlational Analysis

If we assume that there is a natural ordering of technical quality where more freezings mean lower technical quality, then the order of sMOS ratings should be I0>I1>I2>I3>I4. Thus there should be a correlation between technical quality ratings and the expected ordering of integrity impairments in terms of technical quality and the higher that correlation the more reliable a participant would appear to be. Note that this a model-based prediction, as distinct from using the correlation between individual participant scores for impairments and corresponding mean scores for the entire sample. Note that the ordering of impairments and failures was randomized between participants, so that factors such as the peak-end effect [11] should not have systematically affected the aggregated results across participants.

Removing those participants who show systematic violation of the natural ordering of integrity impairments is likely better than removing people because of one or two apparently bad ratings. One or two anomalous data points might have been due to a lapse of concentration or to a temporary misunderstanding. On the other hand, if a person's ratings systematically violate the natural ordering of impairments, then there is evidence for a more global problem with that person's ratings. Consider I0, and the integrity impairments I1, I2, I3, and I4. Disregarding possible contextual factors such as the peak-end effect, or the confounding effect of content quality, we expect that the ordering of corresponding sMOS ratings should be I0>I1>I2>I3>I4. Thus if we assigned the values 5, 4, 3, 2, and one to I0, I1, I2, I3, and I4 respectively, then there should be a strong positive correlation between sMOS scores and those values. In this case the corresponding correlation serves as a measure of how well the person's sMOS scores match the expected ordering.

Fig. 3 shows the resulting correlations between sMOS and the ordered integrity values for the 20 participants (ordered from lowest to higher correlation) who completed the experiment. It can be seen that, using this criterion, six of the participants do comparatively poorly with correlations around 0.4 or under. Most of the participants have correlations in the range .55 to .80, but two of the participants have very high correlations (close to .9).

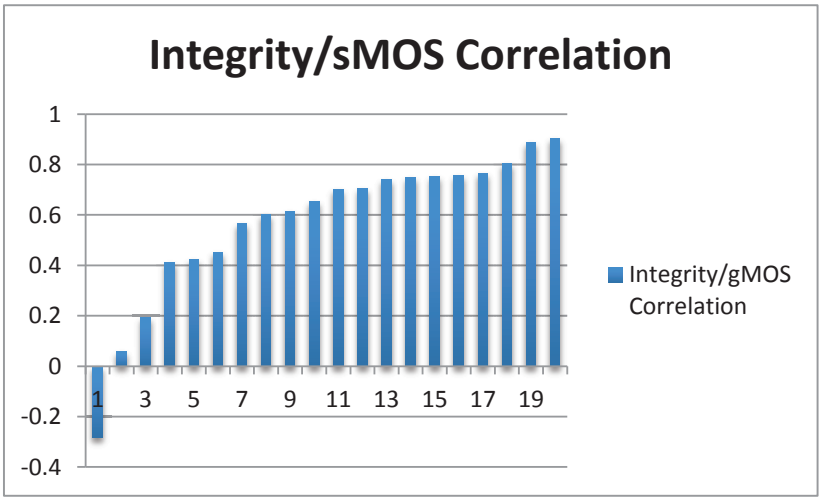

Fig. 3. Correlation between Integrity and sMOS.

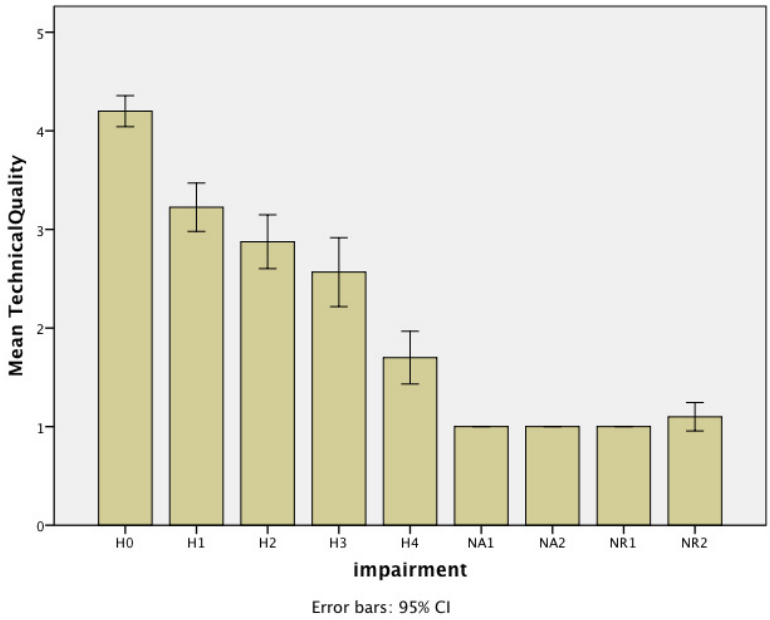

Fig. 4. Revised Mean sMOS Ratings for Participants showing correlations of greater than .7 with the expected ordering of impairments.

4.3. Revised sMOS based on sensitivity to Integrity Impairment

In an attempt to develop a more accurate estimate of sMOS across the impairments we took the ten participants who had correlations of .7 or better between the integrity values and their sMOS scores. We then calculated mean sMOS scores for each of the impairments within that pool of 10 participants. The resulting bar chart is shown in Fig. 4. It can be seen that this estimation of the sMOS values has a number of advantages over the values averaged across all 20 participants in the experiment. First, there is a smooth drop in the integrity values as the level of impairment (number of freezings) increases. Second, non-retainability is judged more harshly with the ratings being anchored at, or very close, to the lower end of the scale. One feature of this interpretation of the sMOS values is that there is a large drop of one point on the rating scale from I0 to I1, and another large drop of close to one point between H3 and H4.

We also tested the impact of content ratings on judged technical quality (sMOS) and overall experience for these ten participants. The correlation between content rating and sMOS was lower for this group (r =0.453 vs. r = 0.522 for that same correlation across all 20 participants). Conversely, the correlation between technical content ratings and overall experience ratings was higher (r = 0.880 vs. r = 0.859). Thus it seems that the screening method used did in fact identify a subgroup of participants that were somewhat more sensitive to judging the technical aspects of the video rather than the content.

4.4. How much impact does content have on sMOS and Overall Experience?

Since a correlation was found between content quality and technical quality we also examined what happened when data was discarded for those trials where the participant gave the highest content rating of 5 to a video.

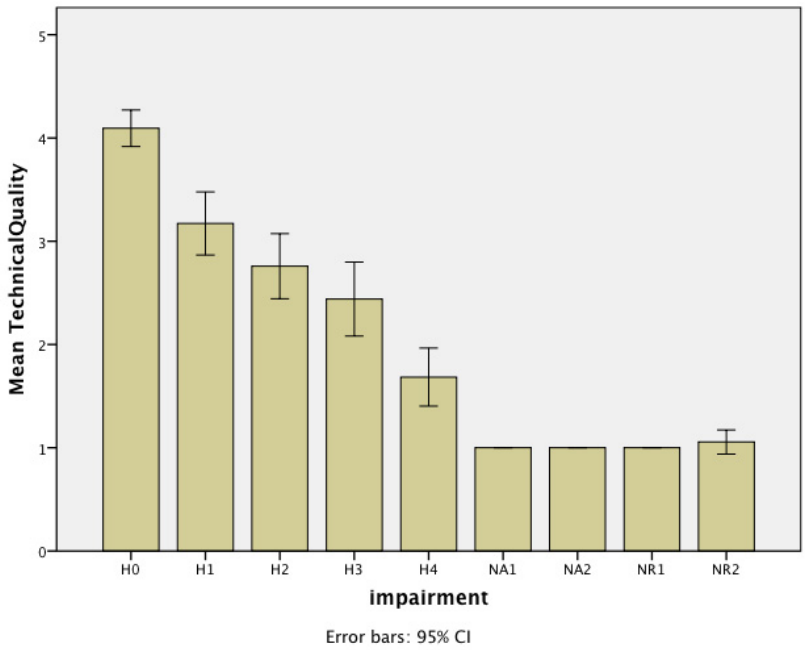

Fig. 5. sMOS.values across impairments for "reliable" participants (videos with the highest content ratings(5) were removed from consideration in constructing this figure).

Fig.5 shows mean sMOS values across impairments across the ten "reliable" participants, where trials with content ratings of 5 have been removed. While the differences between the earlier "reliable" participants only chart and this one (where the highest content ratings have been removed) are only slight, there is perhaps a slightly greater differentiation between the mean sMOS scores for H1, H2, and H3 in this case.

4.5. Do different people judge impairments differently?

We used clustering (K-means analysis as implemented in SPSS) to see whether different groups of people responded to impairments differently in their sMOS ratings. We found the two-cluster solution to provide the best interpretation of the data, placing 12 participants in cluster 1 and the remainder in cluster 2. Differences between the clusters were significant for all impairments except H0, NA1 and NA2. As can be seen in Fig. 6, sMOS ratings for people in cluster 1 were lower for the integrity impairments and for NR, and the pattern of ratings for cluster 1 better matched the expected severities of impairments (sMOS ratings steadily declined with an increasing numbers of freezing events, and NA and NR impairments had uniformly lower sMOS ratings than integrity impairments). All 10 of the "reliable" participants noted using correlation analysis (with a criterion of $r>.7$) were in cluster 1. It is interesting that two different strategies for segmenting the sample gave very similar answers. Since cluster analysis based on sMOS ratings of impairments makes fewer assumptions than correlating sMOS ratings with the expected ranking of impairments we recommend that cluster analysis be considered in future to segment samples of participants so as to evaluate reliability and bias within different subgroups of participants.

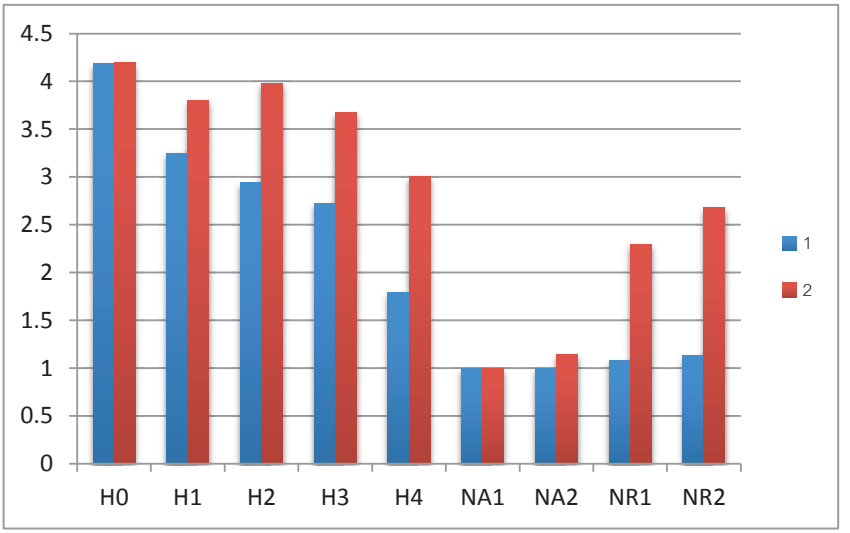

Fig. 6.Participants clustering analysis.

5. Conclusions

While many methods of dealing for screening out unreliable participants based on data properties have been proposed, they typically involve statistical modeling, and assumptions about the distribution of data or relationships within that data. In this paper we demonstrated how a popular method of cluster analysis (K-means analysis) can be used as a distribution free method for characterizing different groups of participants. In the case study reported in this paper, those groups could be interpreted in terms of the degree to which the their respective members were reliable. While clustering of the raw data worked well in the present case, in other situations clustering raw data may not be appropriate if different participants use the rating scale differently. For instance, people who generally assigned higher ratings would tend to be placed in a different cluster than people who assigned lower ratings. Thus in other situations it may be preferable to look at clustering of difference scores between mean impairment ratings, rather than clustering raw data, when screening for unreliable participants.

Acknowledgements

We would like to thank Diba Kaya for assisting with data collection and video content selection for the experiment.

References

1. Malekmohamadi, H., Fernando, W. A. C. & Kondoz, A. M. Automatic QOE Prediction in Stereoscopic Videos. in *2012 IEEE Int. Conf. Multimed. Expo Work.* 581–586 (IEEE, 2012). doi:10.1109/ICMEW.2012.107
2. ITU-T. P.800 : Methods for subjective determination of transmission quality. in *P. 800, Telecommun. Stand. Sect. ITU* (1996).
3. Stevens, S. S. Issues in psychophysical measurement. *Psychol. Rev.* **78,** 426–450 (1971).
4. Lim, J. Hedonic scaling: A review of methods and theory. *Food Qual. Prefer.* **22,** 733–747 (2011).
5. Jaworska, N. & Chupetlovska- Anastasova, A. A review of multidimensional scaling (MDS) and its utility in various psychological domains. *Tutorials Quant. ...* **1,** 1–10 (2009).
6. Sebastian Möller, A. R. *Quality of Experience. Qual. Exp. Adv. Concepts,* (Springer International Publishing, 2014). doi:10.1007/978-3-319-02681-7
7. Leon-Garcia, A. & Zucherman, L. Generalizing MOS to assess technical quality for end-to-end Telecom session. in *2014 IEEE Globecom Work. (GC Wkshps)* 681–687 (IEEE, 2014). doi:10.1109/GLOCOMW.2014.7063511
8. Huynh-Thu, Q., Garcia, M.-N., Speranza, F., Corriveau, P. & Raake, A. Study of Rating Scales for Subjective Quality Assessment of High-Definition Video. *IEEE Trans. Broadcast.* **57,** 1–14 (2011).

9. Hossfeld, T. *et al*. Best Practices for QoE Crowdtesting: QoE Assessment With Crowdsourcing. *IEEE Trans. Multimed.* **16,** 541–558 (2014).
10. ITU-T. P.910 : Subjective video quality assessment methods for multimedia applications. in *Recomm. P. 910, Telecommun. Stand. Sect. ITU* (2008).
11. Daniel Kahneman, A. Tversky (eds, Daniel Kahneman, D. K. Evaluation by Moments: Past and Future.